How Do Humans Distinguish Different People with Identical Names on the Web?: A Cognitive Science Approach

Harumi Murakami
Osaka City University
3-3-138, Sugimoto, Sumiyoshi,
Osaka 558-8585 Japan
+81-6-6605-3375
harumi@media.osaka-cu.ac.jp

Yuki Miyake
Aeon Delight Co., Ltd.
4-1-2, Shonohayama, Suzuka,
Mie 513-0834 Japan
+81-59-375-0666
kakarot2007d@yahoo.co.jp

ABSTRACT
This research investigates how humans distinguish different people with identical names on the web to improve web people search. We asked subjects to classify 20 pages of web people-search results for each of 20 person names and analyzed their decision processes through questionnaire, protocol analysis, and interview. We found that keywords, vocations, works (for a real person, works are those made by the individual and, for a fictional person, works are those in which the individual appears), facial images, and the names of related people are important for distinguishing individuals. We proposed a model for distinguishing individuals and a knowledge-structure model based on the experiment's results.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Selection process

General Terms
Human Factors, Algorithms, Design

Keywords
Web People Search, Person Name Disambiguation, Distinguishing Individual Model, Knowledge-Structure Model

1. INTRODUCTION
The popularity of web people searches continues to rise as the number of people about whom the web provides information increases. Finding information about people on the web is one of the most common activities of Internet users [1]. Person name disambiguation, or distinguishing people with identical names, is becoming more and more important in web searches. There is much research that separates web pages automatically. Currently, the accuracy of person name disambiguation is not sufficient.

We investigate how humans distinguish different people with identical names on the web to improve web people search.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

We describe the method and results of our experiment in Sections 2 and 3. We propose a model for distinguishing individuals and a knowledge-structure model in Section 4. We discuss the significance of our research in Section 5.

2. METHOD
2.1 Overview
We used 14 subjects (9 males and 5 females with an average age of 25).

We collected 400 web pages (HTML files) comprised of 20 web people-search results for each of the 20 Japanese person names used in a related work [2]. At least one of the individual names is famous enough for the authors to identify; however, most are not known to the subjects. We identified which page belongs to which person by referring to Wikipedia, web pages, and so forth. A total of 58 individuals were found in the 400 web pages. We developed a site for the experiment.

We asked the subjects to classify the above 20 web pages for one person name using our developed site. For each person name, we assigned two subjects. We analyzed their decision processes by questionnaire, protocol analysis, and interview.

2.2 Procedure
For each person's name, the procedure of the experiment is as follows.

(1) We first conduct a questionnaire survey to obtain knowledge about the name (whether the subject knows at least one individual with that name).

(2) The subject distinguishes the 20 web people-search results. He or she assigns one or more individual numbers to a web page while speaking out loud about what he or she thinks. The individual for the first result is number one. The procedure is recorded and analyzed by protocol analysis.

(3) After distinguishing pages, the subject answers the questionnaire survey. The two instructions are listed below:

   (a) Rank and write down *ten* or *fewer* terms that are included in either titles, snippets, URLs, or web pages that you think are important for distinguishing individuals. These are called *discriminative keywords*.

   (b) Write *one* term that you think is the most important for characterizing the individuals. You do not have to extract this term from the titles, snippets, URLs, or web pages. These are called *characteristic keywords*.

When the subject participates in an experiment of another person's name, steps (1)-(3) above are repeated. When all experiments are completed for the subject, we conduct a final interview.

3. RESULTS AND ANALYSIS

3.1 Accuracy

We check whether the assigned person number per one page is right to determine accuracy, which was 81% (See Table 1).

Table 1. Number of individuals and accuracy.

Name	No.	Accuracy		Name	No.	Accuracy	
		S1	S2			S1	S2
1	2	0.90	1.00	11	6	0.55	0.25
2	4	1.00	0.90	12	2	0.90	0.90
3	4	0.85	0.85	13	1	1.00	0.65
4	5	1.00	0.75	14	1	1.00	1.00
5	2	0.70	0.65	15	1	1.00	1.00
6	1	0.90	1.00	16	1	1.00	1.00
7	1	0.95	1.00	17	1	0.65	0.95
8	8	0.70	0.55	18	1	0.95	1.00
9	6	0.85	0.65	19	1	0.50	1.00
10	9	0.50	0.50	20	1	0.15	0.90
Total					58	0.81	

Note: Name: person name; No.: number of individuals; S: subject.

3.2 Known and Unknown Names

Among the 40 results (20 person names * two subjects) of the questionnaire, eight results (20%) were for known names.

For real person names, the subjects answered that they did not know these individuals directly, but knew of them through media such as TV. For one fictional person name (a comic book character), a subject answered that he was aware of this individual by reading the comic.

We found that the rate of browsing sites differs between known and unknown names. 35% of subjects browsed sites for known names, while 64% for unknown names.

We think that, for known names, the subjects knew something about at least one individual of that name and they identified that person by looking at the titles or snippets of the page.

3.3 Discriminative and Characteristic Keywords

3.3.1 Overview

We counted the written discriminative and characteristic keywords in the results of our questionnaire survey. The number of discriminative keywords was 329 and that of characteristic keywords was 124.

3.3.2 Example

For example, for the questionnaire result for individual 1 (formally a well-known baseball player, now a sports commentator) for person name 1, the discriminative keywords were *baseball*, *Yomiuri* (name of a baseball team), *monster* (nickname), *comment*, *Kobayashi* (rival person's name), *blank* (known for blank day of contract), *ball*, *Egawaru* (coined-term), and *nine* (number of players on a baseball team). One characteristic keyword for the individual was *baseball*.

3.3.3 Classification of Discriminative and Characteristic Keywords

We classified these terms into eight categories (keywords, vocations, works, related person names, histories, images, URLs, and place names) manually. The categories were set by analyzing data. Although we instructed the subjects to provide descriptive terms, some subjects answered with non-terms, such as photos or URLs. For the former, we created an image category and, for the latter, a URL category. Terms not classified into the seven categories (vocations, works, related person names, histories, images, URLs, and place names), were classified into keywords. Figure 1 shows the result of this classification.

For the discriminative keywords, the top four were keywords (175, 53%), vocations (72, 22%), works (27, 8%), and related person names (24, 7%). 175 keywords were classified into those related to vocations (128, 73%), works (12, 7%), hobbies (8, 5%), sites (6, 3%), personalities (2, 1%), attributes (2, 1%), and others (17, 10%). 72 vocations included 27 organizations and 12 positions.

For the characteristic keywords, the top four were keywords (59, 48%), vocations (44, 35%), works (8, 6%), and related person names (8, 6%). 59 keywords were classified into those related to vocations (35, 59%), hobbies (7, 12%), works (5, 8%), sites (4, 7%), attributes (3, 5%), personalities (2, 3%), and others (3, 5%). 44 vocations included ten organizations and four positions.

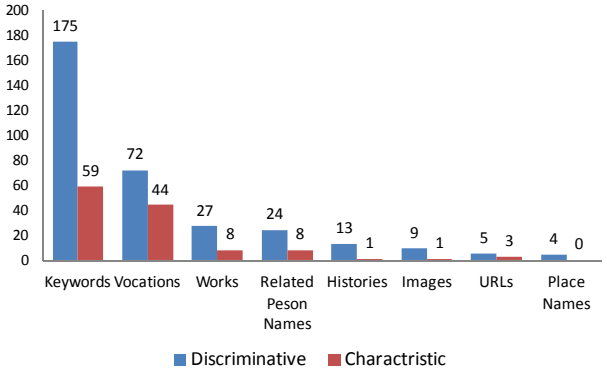
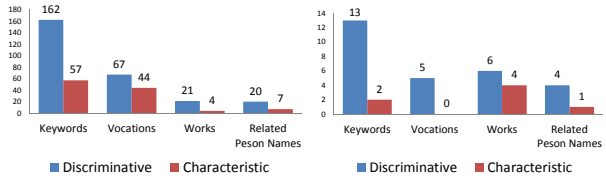


Figure 1. Keywords for overall results.

3.3.4 Classification of Discriminative and Characteristic Keywords of Real or Fictional Individuals

This included 54 real individuals (121 written results) and four fictional individuals (seven written results). We classified the discriminative and characteristic keywords by real and fictional individuals. Figure 2 shows the numbers of keywords, vocations, works, and related person names for the real and fictional individuals.

For real individuals, the works are those made by that individual. For fictional individuals, the works are those in which that individual appears. For real individuals, the related person names are those of other real individuals and, for fictional individuals, the person names are those of other characters in the works in which the fictional individual appears.



(a) Real individuals (b) Fictional individuals

Figure 2. Keywords of top four categories for real and fictional individuals.

For the real individuals, the results are similar to those of the overall results. For both discriminative and characteristic keywords, the top two are keywords and vocations followed by works and related person names. 78% (127/162) and 61% (35/57) were related to vocations for discriminative and characteristic keywords, respectively.

The results for the fictional individuals differed from the real individuals. The top four of the discriminative keywords are keywords, works, vocations, and related person names. The top three characteristic keywords are works, keywords, and related person names with no vocation. 46% (6/13) and 100% (2/2) are related to works for discriminative and characteristic keywords, respectively.

Overall, from our questionnaire investigation, we can conclude that keywords are most important for distinguishing individuals, followed by vocations, then works, and then related person names. When the individual is a real person, information related to the person's vocation(s) is important. When the individual is fictional, information related to works in which the character is included is important.

3.4 Protocol Analysis

As some subjects spoke little during the experiment, we analyzed 20 data (20 person names * 1 subject). We extracted 242 elements and classified them into about the same eight categories as in the previous sections. See Figure 3.

Figure 3. Results of protocol analysis.

The importance of keywords, vocations, and works is the same as in our questionnaire analysis.

We found that checking face images is important (22 out of 23 images) to distinguish individuals, which we did not ascertain through the questionnaire analysis. We also found that checking sites that include the linked pages is important. When there is little information about an individual, for instance, he or she simply posted on news or blog sites, information about the sites is important to distinguish the individual. In addition, when the person is an author and the result is an online book site (e.g., Amazon), these sites provide important clues.

3.5 Interview

Characteristic answers included: "I browsed web sites because there was no knowledge about unknown people to identify individuals." "I imagine something when I distinguish individuals." "It is important to identify individuals whether I get information about faces, vocations, personal histories, achievement, and organizations." "I can identify people when I find one notable achievement." "Facial images are important because I could obtain much information about the person."

3.6 Features of People Who Are Easy to Identify

From the experimental results of unknown person names, we analyzed the features of individuals who are easy to identify. We identified eight features and classified them into four categories (facial images, content, quantity of information, and ease of separation).

(1) Facial images

- (a) There are facial images on the web sites in the top three results for the person name (web pages).

(2) Contents

- (b) Vocations (including organizations) or personal histories are written in snippets or listed on web sites.
- (c) Achievements are listed on web sites.
- (d) Keywords about vocations, personal histories, or achievements are located in titles, snippets, or web sites.
- (e) There is a Wikipedia page.

(3) Quantity of information

- (f) There are more than two search results.

(4) Ease of separation

- (g) There are less than six individuals who share the same name.
- (h) There is no individual who is similar in vocations or fields.

Here, we define personal histories as events, including yearly information such as birth date.

4. COGNITIVE MODELS

Based on the experiments, we propose a model for distinguishing individuals (Figure 4) and a knowledge-structure model (Figure 5).

4.1 A Model for Distinguishing Individuals

The human tendency is to check a people search list from the top down. A person will observe the title, snippet, or URLs of the list and judge whether he can identify the individual using a knowledge-structure. When he can identify the individual, he separates the individual. When he cannot identify the individual or he wants to check the content further, he browses the web site. On browsing the site, he obtains information about the individual and a knowledge-structure for the individual is created or modified. He sometimes selects linked web pages to obtain more information. After that, he separates the individual by checking the knowledge-structure models for individuals (when he has already created other knowledge-structure models).

The above process is repeated until the list is processed.

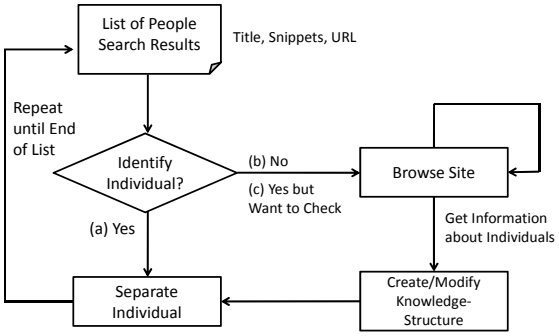


Figure 4. A model for distinguishing individuals.

4.2 A Knowledge-Structure Model

A knowledge-structure model is knowledge about the individual that separates him. A knowledge-structure model is divided into two parts: knowledge about the individual and the site.

Knowledge about the individual has two parts: facial images and text content. Text contents are divided into real and fictional individuals constructed from knowledge about the individual and relationships with other individuals. The knowledge on a real individual includes keywords, vocations, works, and personal histories, while the knowledge on a fictional individual includes keywords and vocations.

Knowledge about a site includes structure of the sites, URLs, and linked pages.

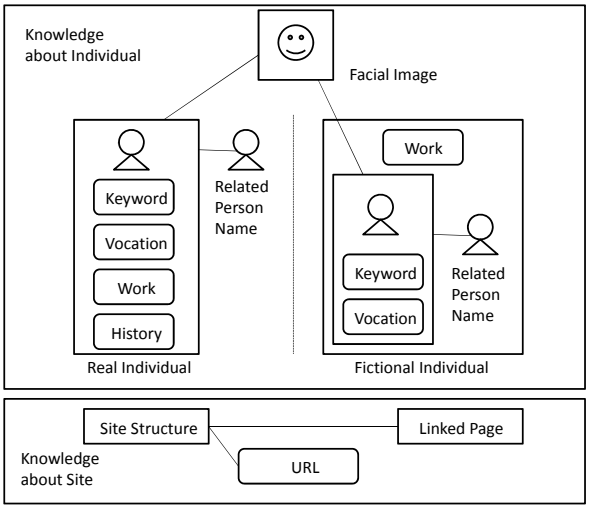


Figure 5. A knowledge-structure model.

5. RELATED WORK AND DISCUSSION

The WePS-2 conducted a competitive evaluation on person attribute extraction on web pages [3]. As named entity recognition (NER) is used in most approaches, Artiles et al. investigated which document features contribute to person name disambiguation and reported that NER only makes a small contribution [4].

There is research that assigns labels to distinguish people. Wan et al. assigned titles (similar to vocations in this research) [5], Ueda et al. assigned vocation-related information including vocations, organizations, and works [6], and Mori et al. assigned keywords to person clusters [7].

The following are our paper's main contributions. First, we investigated how humans identify different people with identical names on the web to obtain knowledge that is helpful for people search. We found that keywords, vocations, works (for an actual person, works are those made by the individual and, for a fictional person, works are those in which the individual appears), facial images, and the names of related people are important for distinguishing individuals. Second, we proposed a model for distinguishing individuals and a knowledge-structure model based on the experiment's results.

We think our findings are useful (a) to consider person name disambiguation algorithms, (b) to develop a person search user interface, and (c) to develop a personal web to be distinguished from other individuals.

6. CONCLUSIONS

We investigated how humans distinguish different people with identical names on the web to improve web people search. We asked subjects to classify 20 pages of web people-search results for each of 20 person names and analyzed their decision processes through questionnaire, protocol analysis, and interview. We found that keywords, vocations, works, facial images, and the names of related people are important for distinguishing individuals. We proposed a model for distinguishing individuals and a knowledge-structure model based on the experiment's results.

7. ACKNOWLEDMENTS

This work was supported by KAKENHI (22500219).

8. REFERENCES

[1] Artiles, J., Gonzalo, J., and Sekine, S. 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *SemEval-2007 Proceedings of the Fourth International Workshop on Semantic Evaluations.* 64-69.

[2] Sato, S., Kazama, K., Fukuda, K., and Murakami, K. 2005. Distinguishing between People on the Web with the Same First and Last Name by Real-world Oriented Web Mining. *IPSJ Transactions on Databases*. 26, 26-36, in Japanese.

[3] Artiles, J., Gonzalo, J. and Sekine, S. Artiles. 2009. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task J. In *2nd Web People Search Evaluation Workshop WePS 2009, 18th WWW Conference*.

[4] Artiles, J., Amigo, E, and Gonzalo, J. 2009. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* 534-542.

[5] Wan, X., Gao, J., Li, M., and Ding, B. 2005. Person Resolution in Person Search Results: WebHawk. In *Proceedings of the 14th ACM International conference on Information and knowledge management*. 163-170.

[6] Ueda, H., Murakami, H., and Tatsumi, S. 2009. Assigning Vocation-Related Information to Person Clusters for Web People Search Results. In *Proceedings of the 2009 Global Congress on Intelligent Systems*. 4, 248-253.

[7] Mori, J., Matsuo, Y., and Ishizuka, M. 2005. Personal Keyword Extraction from the Web. *Journal of Japanese Society for Artificial Intelligence*. 20, 337-345, in Japanese.